On the combinatorics of sparsification

Fenix W.D. Huang[1] and Christian M. Reidys*[1]

[1]Department of Mathematic and Computer science, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

Email: Fenix W.D. Huang - fenixprotoss@gmail.com; Christian M. Reidys*- duck@santafe.edu;

*Corresponding author

Abstract

Background: We study the sparsification of dynamic programming folding algorithms of RNA structures. Sparsification applies to the mfe-folding of RNA structures and can lead to a significant reduction of time complexity.

Results: We analyze the sparsification of a particular decomposition rule, Λ^*, that splits an interval for RNA secondary and pseudoknot structures of fixed topological genus. Essential for quantifying the sparsification is the size of its so called candidate set. We present a combinatorial framework which allows by means of probabilities of irreducible substructures to obtain the expected size of the set of Λ^*-candidates. We compute these expectations for arc-based energy models via energy-filtered generating functions (GF) for RNA secondary structures as well as RNA pseudoknot structures. For RNA secondary structures we also consider a simplified loop-energy model. This combinatorial analysis is then compared to the expected number of Λ^*-candidates obtained from folding mfe-structures. In case of the mfe-folding of RNA secondary structures with a simplified loop energy model our results imply that sparsification provides a reduction of time complexity by a constant factor of 91% (theory) versus a 96% reduction (experiment). For the "full" loop-energy model there is a reduction of 98% (experiment).

Conclusions: Our result show that the polymer-zeta property, describing the probability of an irreducible structure over an interval of length m does not hold for RNA structures. As a result sparsification of the Λ^*-decomposition rule does not lead to a linear reduction of the set of candidates. We show that under general assumptions the expected number of Λ^*-candidates is $\Theta(n^2)$, the constant reduction being in the range of 95%. The sparsification of the Λ^*-decomposition rule for RNA pseudoknotted structures of genus 1 leads to an expected number of candidates of $\Theta(n^2)$. The effect of sparsification is sensitive to the employed energy model.

Background

An RNA sequence is a linear, oriented sequence of the nucleotides (bases) **A,U,G,C**. These sequences "fold" by establishing bonds between pairs of nucleotides. Bonds cannot form arbitrarily: a nucleotide can at most establish one Watson-Crick base pair **A-U** or **G-C** or a wobble base pair **U-G**, and the global conformation of an RNA molecule is determined by topological constraints encoded at the level of secondary structure, i.e., by the mutual arrangements of the base pairs [1].

Secondary structures can be interpreted as (partial) matchings in a graph of permissible base pairs [2]. They can be represented as diagrams, i.e. graphs over the vertices $1, \ldots, n$, drawn on a horizontal line with bonds (arcs) in the upper halfplane. In this representation one refers to a secondary structure without crossing arcs as a *simple* secondary structure and pseudoknot structure, otherwise, see Figure 1.

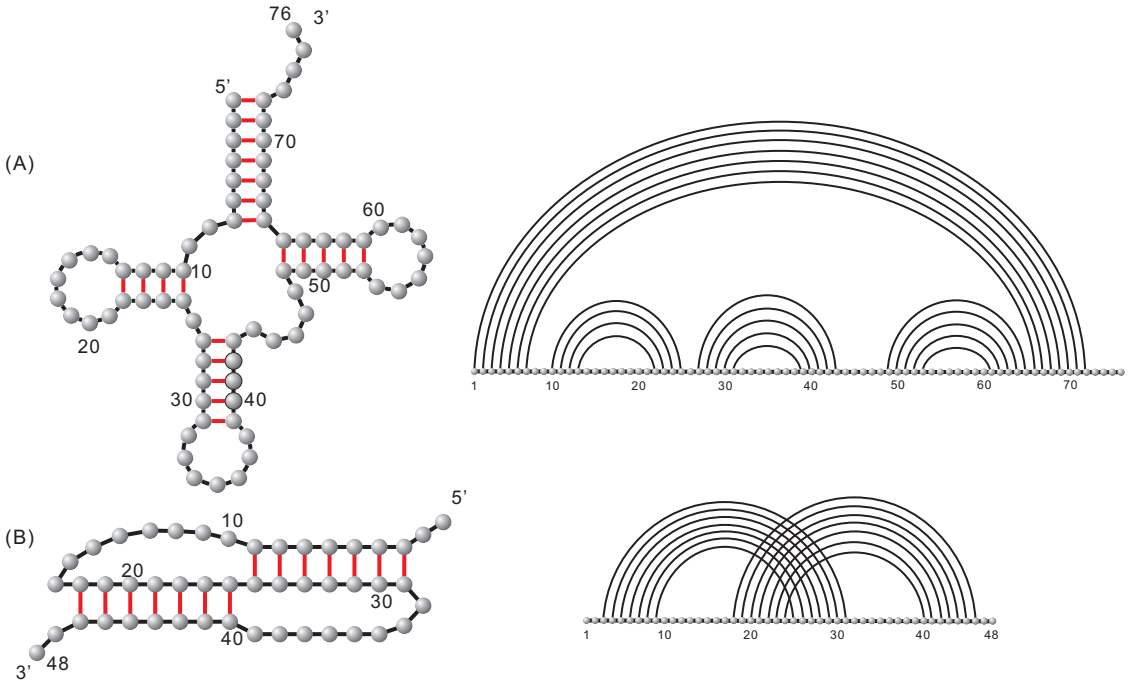

Figure 1: RNA structures as planar graphs and diagrams. (A) an RNA secondary structure and (B) an RNA pseudoknot structure.

Folded configurations are energetically somewhat optimal. Here energy means free energy, which is dominated by the loops forming between adjacent base pairs and not by the hydrogen bonds of the individual base pairs [3]. In addition sterical constraints imply certain minimum arc-length conditions for minimum

free energy configurations [4]. In particular, only configurations without isolated bonds and without bonds of length one (formed by immediately subsequent nucleotides) are observed in RNA structures. In this paper, optimize a problem we meas maximize the score but not to minimize the free energy.

For a given RNA sequence polynomial-time dynamic programming (DP) algorithms can be devised, finding such minimal energy configurations. The most commonly used tools predicting simple RNA secondary structure `mfold` [5] and the `Vienna RNA Package` [6], are running at $O(N^2)$ space and $O(N^3)$ time solution. In the following we omit "simple" and refer to secondary structures containing crossing arcs as pseudoknot structures.

Generalizing the matrices of the DP-routines of secondary structure folding [5,6] to gap-matrices [7], leads to a DP-folding of pseudoknotted structures [7] (`pknot-R&E`) with $O(n^4)$ space an $O(n^6)$ time complexity. The following references provide a certainly incomplete list of DP-approaches to RNA pseudoknot structure prediction using various structure classes characterized in terms of recursion equations and/or stochastic grammars: [7–19]. The most efficient algorithm for pseudoknot structures is [14] (`pknotsRG`) having $O(n^2)$ space and $O(n^4)$ time complexity. This algorithm however considers only a few types of pseudoknots.

RNA secondary structures are exactly structures of topological genus zero [20]. The topological classification of RNA structures [21–23] has recently been translated into an efficient DP algorithm [19]. Fixing the topological genus of RNA structures implies that there are only finitely many types, the so called irreducible shadows [23].

Sparsification is a method tailored to speed up DP-algorithms predicting mfe-secondary structures [24,25]. The idea is to prune certain computation paths encountered in the DP-recursions, see Figure 2. To make the key point, let us consider the case of RNA secondary structure folding. Here sparsification reduces the DP-recursion paths to be based on so called candidates. A candidate is in this case an interval, for which the optimal solution cannot be written as a sum of optimal solutions of sub-intervals, see Figure 3. Tracing back these candidates gives rise to "irreducible" structures and the crucial observation is here that these irreducibles appear only at a low rate. This means that there are only relatively few candidates, which in turn implies a significant reduction in time and space complexity.

Sparsification has been applied in the context of RNA-RNA interaction structures [26] as well as RNA pseudoknot structures [27]. In difference to RNA secondary structures, however, not every decomposition rule in the DP-folding of RNA pseudoknot structures is amendable to sparsification. By construction, sparsification can only be applied for calculating mfe-energy structures. Since the computation of the partition function [12, 28] needs to take into account *all* sub-structures, sparsification does not work.

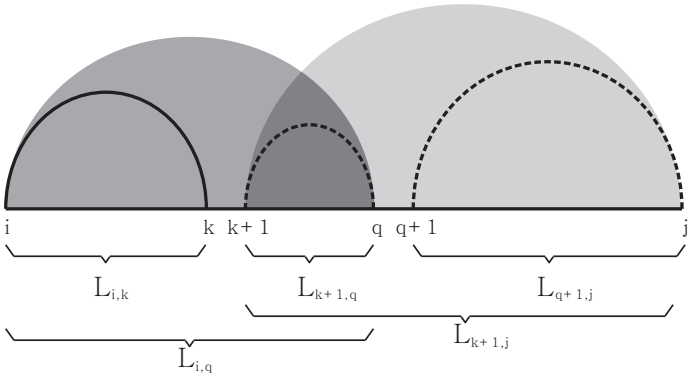

Figure 2: Sparsification of secondary structure folding. Suppose the optimal solution $L_{i,j}$ is obtained from the optimal solutions $L_{i,k}$, $L_{k+1,q}$ and $L_{q+1,j}$. Based on the recursions of the secondary structures, $L_{i,k}$ and $L_{k+1,q}$ produce an optimal solution of $L_{i,q}$. Similarly, $L_{k+1,q}$ and $L_{q+1,j}$ produce an optimal solution of $L_{k+1,j}$. Now, in order to obtain an optimal solution of $L_{i,j}$ it is sufficient to consider either the grouping $L_{i,q}$ and $L_{q+1,j}$ or $L_{i,k}$ and $L_{k+1,j}$.

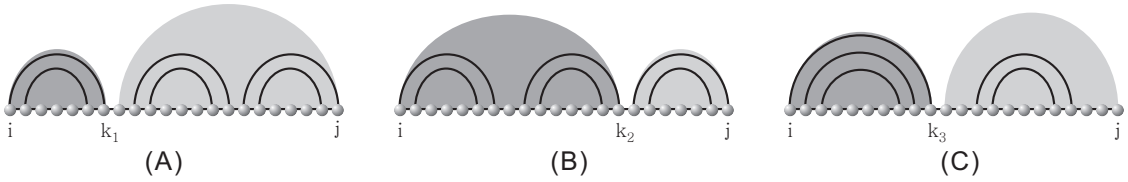

Figure 3: What sparsification can and cannot prune: (A) and (B) are two computation paths yielding the same optimal solution. Sparsification reduces the computation to path (A) where S_{i,k_1} is irreducible. (C) is another computation path with distinct leftmost irreducible over a different interval, hence representing a new candidate that cannot be reduced to (A) by the sparsification.

For the mfe-folding of RNA secondary structures considerable attention has been paid in order to validate that the set of candidates is small. The idea here is that irreducibles are contained in short, "rainbow"-like arcs. To be precise, the gain is $O(n)$, if secondary structure satisfy the so called *polymer-zeta property* [29,30]: The latter quantifies the probability of an arc of length m to be $\leq b\ m^{-c}$, where $b > 0$, $c > 1$. Note that these arcs confine in case of secondary structures irreducible structures, that is arcs and irreducibility are tightly connected.

In pseudoknotted RNA structures however, we have crossing arcs and the associated notion of irreducible structures differs significantly from that of RNA secondary structures. The polymer-zeta property is theoretically justified by means of modeling the 2D folding of a polymer chain as a self-avoiding walk (SAW) in a 2D lattice [31]. More evidence of the polymer-zeta property for RNA secondary structures has been collected via the NCBI database [32] of mfe-RNA structures.

In this paper we study the sparsification of the decomposition rule Λ^* that splices an interval $[25, 27]$ in

the context of the DP-folding of RNA pseudoknot structures of fixed topological genus. Our paper provides a combinatorial framework to quantify the effects of sparsifying the Λ^*-decomposition rule.

We shall prove that the candidate set [24, 25, 27] is indeed small. Our argument is based on assuming a specific distribution of irreducible structures within mfe-structures. Namely we assume these irreducibles to appear with probability $\mathbf{f}^*(n,j)/\mathbf{f}(n,j)$, where we assume e to be a fixed parameter and $\mathbf{F}(z,e) = \sum \mathbf{f}_{n,j} z^n e^j$ to be a bivariate (energy-filtered) generating function whose associated generation function of irreducibles is $\mathbf{F}^*(z,e) = \sum \mathbf{f}^*_{n,j} z^n e^j$.

While this energy-filtration seems to be reparameterization of the notion of "stickiness" [33], it is really fundamentally different. This becomes clear when considering loop-based energies which distinguishes energy and arcs. Clearly when folding random sequences one weights the latter around 6/16, reminiscent of the probability of two given positions to be compatible. The energy however is fairly independent as it really depends on the particular loop-type.

We obtain these energy-filtered GFs also for RNA pseudoknot structures of fixed topological genus. This provides new insights into the improvements of the sparsification of the concatenation-rule Λ^* in the presence of cross serial interactions. Our observations complement the detailed analysis of Backofen [25,27]. We show that although for pseudoknot structures of fixed topological genus [22, 23] the effect of sparsification on the global time complexity is still unclear, the decomposition rule that splits an interval can be sped up significantly.

Sparsification

The general idea of sparsification [24,25,27] is following: let $V = \{v_1, v_2, \ldots\}$ be a set whose elements v_i are unions of pairwise disjoint intervals. Let furthermore L_v denote an optimal solution (a positive number or score) of the DP-routine over v. By assumption L_v is recursively obtained. Suppose the optimal solution L_v is given by $L_v = L_{v_1} + L_{v_2} + L_{v_3}$, where $v = v_1 \dot\cup v_2 \dot\cup v_3$. Then, under certain circumstances, the DP-routine may interpret L_v either as $(L_{v_1} + L_{v_2}) + L_{v_3}$ or as $L_{v_1} + (L_{v_2} + L_{v_3})$, see Figure 4. To be precise, this situation is encountered iff

- there exists an optimal solution $L_{v_1'}$ for a sub-structure over v_1' where $v_1' = v_1 \dot\cup v_2$ via Λ_2 and L_v is obtained from $L_{v_1'}$ and L_{v_3} via Λ_1,
- there exists an optimal solution $L_{v_2'}$ for a sub-structure over v_2' where $v_2' = v_2 \dot\cup v_3$ via Λ_3 and L_v is obtained by L_{v_1} and $L_{v_2'}$ via Λ_1.

Given a decomposition

$$L_v = \underbrace{\underbrace{L_{v_1} + L_{v_2}}_{\Lambda_2} + L_{v_3}}_{\Lambda_1},$$

we call Λ_2 s-compatible to Λ_1 if there exists a decomposition rule Λ_3 such that

$$L_v = \underbrace{L_{v_1} + \underbrace{L_{v_2} + L_{v_3}}_{\Lambda_3}}_{\Lambda_1}.$$

Note that if Λ_2 is s-compatible to Λ_1 then Λ_3 is s-compatible to Λ_1. To summarize

Definition 1. (s-compatible) Suppose L_v is the optimal solution for S_v over v, $L_v = L_{v_1'} + L_{v_3}$ under decomposition rule Λ_1. L_{v_1} is obtained from two optimal solutions L_{v_1} and L_{v_2} under rule Λ_2. Then Λ_2 is called *s-compatible* to Λ_1 if there exist some rule Λ_3 such that $L_{v_2'} = L_{v_2} + L_{v_3}$ and $L_v = L_{v_1} + L_{v_2'}$.

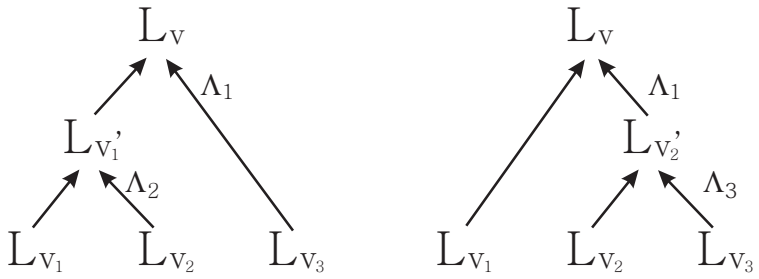

Figure 4: Sparsification: L_v is alternatively realized via L_{v_1} and $L_{v_2'}$, or $L_{v_1'}$ and L_{v_3}. Thus it is sufficient to only consider one of the computation paths.

Figure 4 depicts two such ways that realize the same optimal solution L_v. Sparsification prunes any such multiple computations of the same optimal value.

We next come to the important concept of candidates. The latter mark the essential computation paths for the DP-routine.

Definition 2. (Candidates) Suppose L_v is an optimal solution. We call v is a Λ-*candidate* if for any $v_1 \subsetneq v$ obtained by Λ and $v = v_1 \dot{\cup} v_2$, we have

$$L_v > L_{v_1} + L_{v_2}$$

and we shall denote the set of Λ-candidates set by Q^{Λ}.

Lemma 1. *[24, 27] Suppose Λ_2 is s-compatible to Λ_1 then any optimal solution L_v can be obtained via Λ_2-candidates.*

By construction a Λ_2-candidate v is a union of disjoint intervals such that its optimal solution L_v cannot be obtained via a Λ_2-splitting. This optimal solution allows to construct a non-unique arc-configuration (sub-structure) over v [5,6] and the above Λ_2-splitting consequently translates into a splitting of this sub-structure. This connects the notion of Λ_2-candidates with that of sub-structures and shows that a Λ_2-candidate implies an sub-structure that is Λ_2-irreducible.

In the case of sparsification of RNA secondary structures we have one basic decomposition rule Λ^* acting on intervals, namely Λ^* splices an interval into two disjoint, subsequent intervals. The implied notion of a Λ^*-irreducible sub-structure is that of a sub-structure nested in an maximal arc, where maximal refers to the partial order $(i,j) \leq (i',j')$ iff $i' \leq i \ \wedge j \leq j'$. This observation relates irreducibility to that or arcs and following this line of thought [24] identifies a specific property of polymer-chains introduced in [29,30] to be of relevance for the size of candidate sets:

Definition 3. (Polymer-zeta property) Let $P(i,j)$ denotes the probability of a structure over an interval $[i,j]$ under some decomposition rule Λ. Then we say Λ follows the polymer-zeta property if $P(i,j) = bm^{-c}$ for some constant $b, c > 0$.

This property is theoretically justified by means of modeling the 2D folding of a polymer chain as a self-avoiding walk (SAW) in a 2D lattice [31].

RNA secondary structures

In this section we recall some results of [24,25] on the sparsification of RNA secondary structures. Secondary structure satisfies a simple recursion which gives the optimal solution over $[i,j]$ by $L_{i,j} = \max\{V_{i,j}, W_{i,j}\}$, where $V_{i,j}$ denotes the optimal solution in which (i,j) is a base pair, and $W_{i,j}$ denotes the optimal solution obtained by adding the optimal solutions of two subsequent intervals, respectively. Note that the optimal solution over a single vertex is denoted by $L_{i,i}$. We have the recursion equation for $V_{i,j}$ and $W_{i,j}$:

$$\begin{aligned} (\Lambda_1) \quad V_{i,j} &= L_{i+1,j-1} + f(i,j), \\ (\Lambda_2) \quad W_{i,j} &= \max_{i<k<j}\{L_{i,k} + L_{k+1,j}\}, \end{aligned}$$

where $f(i,j)$ is the score when (i,j) form a base pair, see Figure. 5. In case two positions, i,j in the sequence are incompatible then we have $f(i,j) = -\infty$.

An interval $[i,j]$ is a Λ^*-candidate if the optimal solution over $[i,j]$ is given by $L_{i,j} = V_{i,j} > W_{i,j}$. Indeed, $[i,j]$ is a candidate iff $[i,j]$ is in the candidate set of Λ^*, and we denote the set Q^{Λ^*} by Q. Suppose the

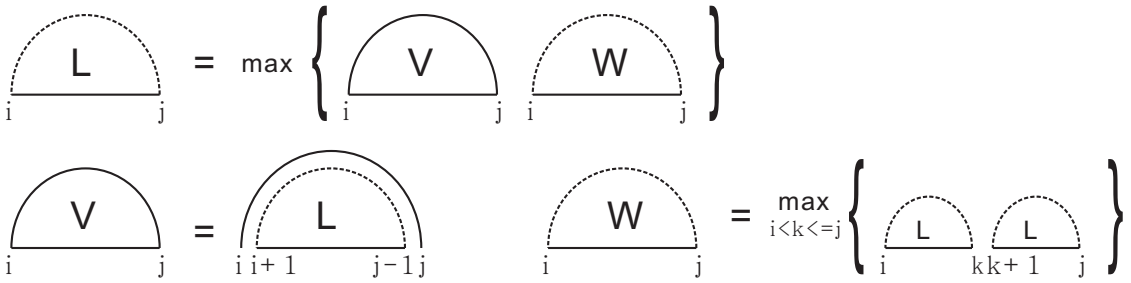

Figure 5: The recursion solving the optimal solution for secondary structures.

optimal solution $W_{i,j}$ is given by $W_{i,j} = L_{i,q} + L_{q+1,j}$ and suppose we have $L_{i,q} = L_{i,k} + L_{k+1}$. Then since $[i, q]$ is not a candidate, Lemma 1 shows that we can compute $W_{i,j} = L_{i,k} + L_{k+1,j}$, where $[i, k]$ is a candidate.

Accordingly, the recursion for $W_{i,j}$ can be based on candidates, i.e. $W_{i,j} = \max_{[i,k]\in Q}\{L_{i,k} + L_{k+1,j}\}$. Clearly, the bottleneck for computing the recursion is the calculation of $W_{i,j}$, which requires $O(n^3)$ time. Applying sparsification, this recursion is based on candidates $[i, k]$. Suppose we have Z such candidates, then the time complexity reduces to $O(nZ)$, since the optimal solution is necessarily based on a candidate. Once the latter is identified the expression $L_{k+1,j}$ requires only $O(n)$ time complexity. In the worst case, Q contains $O(n^2)$ elements.

The polymer-zeta property however implies that the expectation of Z is given by $\sum_{i\geq 1}^{n}\sum_{j=i} b(j-i)^{-c}$ where b and c are constants and $c > 1$. We can conclude from the polymer-zeta property that $Z = O(n)$ and accordingly the runtime reduces to $O(n) \cdot O(n) = O(n^2)$.

RNA pseudoknot structures

Sparsification can also be applied to the DP-algorithm folding RNA structures with pseudoknots [27]. In contrast to the decomposition rule Λ^* that spliced an interval into two subsequent intervals, we encounter in the grammar for pseudoknotted structures additional more complex decomposition rules [7]. As shown in [27] there exist some decomposition rules which are not s-compatible and which can accordingly not be sparsified at all, see Figure 6. For instance, given a decomposition rule Λ in `pknot-R&E` subsequent decomposition rules which are s-compatible to Λ are referred to as split type of Λ [27].

In the following we will study RNA pseudoknot structures of fixed topological genus, see Section **Diagrams, surfaces and some generating functions** for details. An algorithm folding such pseudoknot structures, `gfold`, has been presented in [19]. The decomposition rules that appear in `gfold` are reminiscent to those of `pknot-R&E` but as they restrict the genus of sub-structures the iteration of gap-matrices is severely

restricted and the effect of sparsification of these decompositions is significantly smaller.

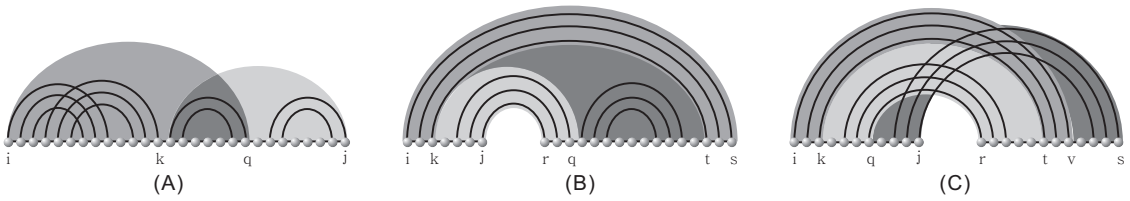

Figure 6: Decomposition rules for pseudoknot structures of fixed genus. (A) three decompositions via the rule Λ^*, which is s-compatible to itself. We show that for Λ^* we obtain a linear reduction in time complexity. (B) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are s-compatible to Λ_1. A quantification of the candidate set is not implied by the polymer-zeta property. (C) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are not s-compatible to Λ_1.

In the following, we restrict our analysis to the decomposition rule Λ^* which splices an interval into two subsequent intervals. Expressed in combinatorial language, Λ^* cuts the backbone of an RNA pseudoknot structure of fixed genus g over one interval without cutting a bond.

Methods

Diagrams and genus filtration

In this section we recall some facts about diagrams and pass from diagrams to surfaces in order to be able to formulate what we mean by an RNA pseudoknot structure of fixed genus g. Most of this section is derived from [23, 34] with the exception of Lemma 2 and Theorem 2, which are new and key for the subsequent analysis of Λ^*-candidates.

A diagram is a labeled graph over the vertex set $[n] = \{1, \ldots, n\}$ in which each vertex has degree ≤ 3, represented by drawing its vertices in a horizontal line. The backbone of a diagram is the sequence of consecutive integers $(1, \ldots, n)$ together with the edges $\{\{i, i+1\} \mid 1 \leq i \leq n-1\}$. The arcs of a diagram, (i, j), where $i < j$, are drawn in the upper half-plane. We shall distinguish the backbone edge $\{i, i+1\}$ from the arc $(i, i+1)$, which we refer to as a 1-arc. A stack of length ℓ is a maximal sequence of "parallel" arcs, $((i, j), (i+1, j-1), \ldots, (i+(\ell-1), j-(\ell-1)))$ and is also referred to as a ℓ-stack, see Figure 7.

We shall consider diagrams as fatgraphs, $\mathbb{G}$, that is graphs G together with a collection of cyclic orderings, called fattenings, one such ordering on the half-edges incident on each vertex. Each fatgraph $\mathbb{G}$ determines an oriented surface $F(\mathbb{G})$ [35,36] which is connected if G is and has some associated genus $g(G) \geq 0$ and number $r(G) \geq 1$ of boundary components. Clearly, $F(\mathbb{G})$ contains G as a deformation retract [37]. Fatgraphs were first applied to RNA secondary structures in [38] and [39].

A diagram $\mathbb{G}$ hence determines a unique surface $F(\mathbb{G})$ (with boundary). Filling the boundary components

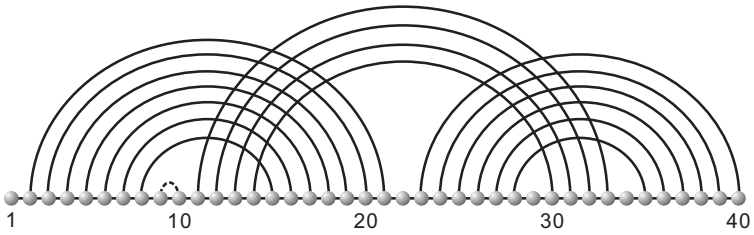

Figure 7: RNA structures and diagram representation. A diagram over $\{1,\ldots,40\}$. The arcs $(1,21)$ and $(11,33)$ are crossing and the dashed arc $(9,10)$ is a 1-arc which is not allowed. This structure contains 3 stacks with length 7, 4 and 6, from left to right respectively.

with discs we can pass from $F(\mathbb{G})$ to a surface without boundary. Euler characteristic, χ, and genus, g, of this surface is given by $\chi = v - e + r$ and $g = 1 - \frac{1}{2}\chi$, respectively, where v, e, r is the number of discs, ribbons and boundary components in $\mathbb{G}$, [37]. The genus of a diagram is that of its associated surface without boundary and a diagram of genus g is referred to as g-diagram.

A g-diagram without arcs of the form $(i, i+1)$ (1-arcs) is called a g-structure. A g-diagram that contains only vertices of degree three, i.e. does not contain any vertices not incident to arcs in the upper halfplane, is called a g-matching. A stack of length τ is a maximal sequence of "parallel" arcs,

$$((i,j),(i+1,j-1),\ldots,(i+\tau,j-\tau)).$$

A diagram is called irreducible, if and only if it cannot be split into two by cutting the backbone without cutting an arc.

Let $\mathbf{c}_g(n)$ and $\mathbf{d}_g(n)$ denote the number of g-matchings and g-structures having n-arcs and n vertices, respectively, with GF

$$\mathbf{C}_g(z) = \sum_{n=0}^{\infty} \mathbf{c}_g(n) z^n \qquad \mathbf{D}_g(z) = \sum_{n=0}^{\infty} \mathbf{d}_g(n) z^n.$$

The GF $\mathbf{C}_g(z)$ has been computed the context of the virtual Euler characteristic of the moduli-space of curves in [34] and $\mathbf{D}_g(z)$ can be derived from $\mathbf{C}_g(z)$ by means of symbolic enumeration [23]. The GF of genus zero diagrams $\mathbf{C}_0(z)$ is wellknown to be the GF of the Catalan numbers, i.e., the numbers of triangulations of a polygon with $(n+2)$ sides,

$$\mathbf{C}_0(z) = \frac{1-\sqrt{1-4z}}{2z}.$$

As for $g \geq 1$ we have the following situation [23]

Theorem 1. *Suppose $g \geq 1$. Then the following assertions hold*

(a) $\mathbf{D}_g(z)$ *is algebraic and*

$$\mathbf{D}_g(z) \quad = \quad \frac{1}{z^2-z+1}\ \mathbf{C}_g\left(\frac{z^2}{(z^2-z+1)^2}\right). \tag{1}$$

In particular, we have for some constant a_g *depending only on* g *and* $\gamma \approx 2.618$*:*

$$[z^n]\mathbf{D}_g(z) \sim a_g\, n^{3(g-\frac{1}{2})}\gamma^n. \tag{2}$$

(b) *the bivariate GF of* g*-structures over* n *vertices, containing exactly* m *arcs,* $\mathbf{E}_g(z,t)$*, is given by*

$$\mathbf{E}_g(z,t) = \frac{1}{tz^2-z+1}\mathbf{D}_g\left(\frac{t\ z^2}{(t\ z^2-z+1)^2}\right). \tag{3}$$

Irreducible g-structures

In the context of Λ^*-candidates we observed that irreducible substructures are of key importance. It is accordingly of relevance to understand the combinatorics of these structures. To this end let $\mathbf{D}^*_g(z) = \sum_{n=0}^{\infty}\mathbf{D}^*_g(n)z^n$ denote the GF of irreducible g-structures.

Lemma 2. *For* $g \geq 0$*, the GF* $\mathbf{D}^*_g(z)$ *satisfies the recursion*

$$\begin{aligned}\mathbf{D}^*_0(z) &= 1-\frac{1}{\mathbf{D}_0(z)}\\ \mathbf{D}^*_g(z) &= -\frac{(\mathbf{D}^*_0(z)-1)\mathbf{D}_g(z)+\sum_{g_1=1}^{g-1}\mathbf{D}^*_{g_1}(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}.\end{aligned}$$

For a proof of Lemma 2, see Section **Proofs**.

Theorem 2. *For* $g \geq 1$ *we have*

(a) *the GF of irreducible* g*-structures over* n *vertices is given by*

$$\mathbf{D}^*_g(z) = (z^2-z+1)\left(\frac{\mathbf{U}_g(u)}{(1-4u)^{3g-\frac{1}{2}}}+\frac{\mathbf{V}_g(u)}{(1-4u)^{3g-1}}\right), \tag{4}$$

where $u = \frac{z^2}{(z^2-z+1)^2}$*,* $\mathbf{U}_g(z)$ *and* $\mathbf{V}_g(z)$ *are both polynomials with lowest degree at least* $2g$*, and* $\mathbf{U}_g(1/4)$*,* $\mathbf{V}_g(1/4) \neq 0$*. In particular, for some constant* $k_g > 0$ *and* $\gamma \approx 2.618$*:*

$$\mathbf{D}^*_g(n) \sim k_g n^{3(g-\frac{1}{2})}\gamma^n. \tag{5}$$

(b) *the bivariate GF of irreducible* g*-structures over* n *vertices, containing exactly* m *arcs,* $\mathbf{E}^*_g(z,t)$*, is given by*

$$\mathbf{E}^*_g(z,t) = (tz^2-z+1)\left(\frac{\mathbf{U}_g(v)}{(1-4v)^{3g-\frac{1}{2}}}+\frac{\mathbf{V}_g(v)}{(1-4v)^{3g-1}}\right), \tag{6}$$

where $v = \frac{tz^2}{(tz^2-z+1)^2}$*.*

We shall postpone the proof of Theorem 2 to Section **Proofs**.

The main result

In Section **Sparsification** we observed that sparsification applies to the decomposition rule Λ^*, which effectively splices off an irreducible sub-structure (diagram). This notion of Λ^*-irreducibility is indeed compatible by the notion of combinatorial irreducibility introduced in Section **Diagrams, surfaces and some generating functions**, see Figure 8. An optimal solution for the original structure is obtained from an optimal solution of the spliced, Λ^*-irreducible, sub-structure and an optimal solution for the remaining sub-structure.

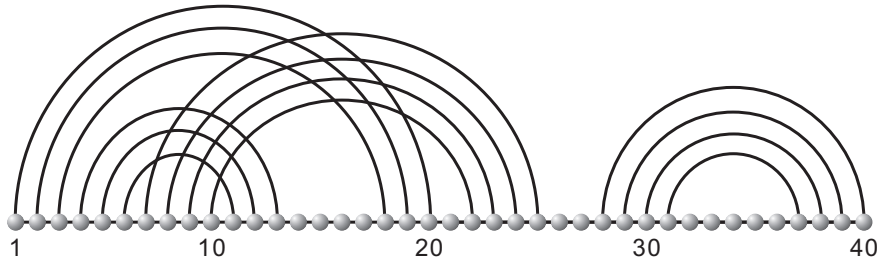

Figure 8: Irreducibility relative to a decomposition rule: the rule Λ^* splitting $S_{i,j}$ to $S_{i,k}$ and $S_{k+1,j}$, $S_{1,40}$ is not Λ^*-irreducible, while $S_{1,25}$ and $S_{28,40}$ are. However, for the decomposition rule Λ_2, which removes the outmost arc, $S_{28,40}$ is not Λ_2-irreducible while $S_{1,25}$ is.

Folded configurations are energetically optimal and dominated by the stacking of adjacent base pairs [3], as well as minimum arc-length conditions [4] discussed before.

In the following we mimic some form of minimum free energy g-structures: inspired by the Nussinov energy model [40] we consider the weight of a g-structure over n vertices to be given by η^ℓ, where ℓ is the number of arcs for some $\eta \geq 1$ [33]. Note that the case $\eta = 1$ corresponds to the uniform distribution, i.e. all g-structure have identical weight.

This approach requires to keep track of the number of arcs, i.e. we need to employ bivariate GF. In Theorem 1 **(b)** we computed this bivariate GF and in Theorem 2 **(b)** we derived from this bivariate GF $\mathbf{E}^*_g(z,t)$, the GF of irreducible g-structures over n vertices containing ℓ arcs.

The idea now is to substitute for the second indeterminant, t, some fixed $\eta \in \mathbb{R}$. This substitution induces the formal power series

$$\mathbf{D}_{g,\eta}(z) = \mathbf{E}_g(z,\eta),$$

which we regard as being parameterized by η. Obviously, setting $\eta = 1$ we recover $\mathbf{D}_g(z)$, i.e. we have $\mathbf{D}_g(z) = \mathbf{D}_{g,1}(z) = \mathbf{E}_g(z,1)$. Note that for $\eta > 1/4$, the polynomial $\eta z^2 - z + 1$ has no real root. Thus we have for $\eta > 1/4$ the asymptotics

$$\mathbf{d}_{g,\eta}(n) \sim a_{g,\eta} n^{3(g-\frac{1}{2})} \gamma_\eta^n \quad \text{and} \quad \mathbf{d}^*_{g,\eta}(n) \sim k_{g,\eta} n^{3(g-\frac{1}{2})} \gamma_\eta^n, \tag{7}$$

with identical exponential growth rates as long as the supercritical paradigm [41] applies, i.e. as long as γ_η, the real root of minimal modulus of

$$\left(\frac{\eta\, z^2}{(\eta\, z^2 - z + 1)^2}\right) = \frac{1}{4},$$

is smaller than any singularity of $\frac{1}{\eta z^2 - z + 1}$. In this situation η affects the constant $a_{g,\eta}$ and the exponential growth rate γ_η but *not* the sub-exponential factor $n^{3(g-\frac{1}{2})}$. The latter stems from the singular expansion of $\mathbf{C}_g(z)$. Analogously, we derive the η-parameterized family of GF $\mathbf{D}^*_{g,\eta}(z) = \mathbf{E}^*_g(z,\eta)$. Assuming a random sequence has on average a probability at most 6/16 to form a base pair we fix in the following $\eta = 6e/16 \approx 1.0125$, where e is the Euler number. By abuse of notation we will omit the subscript η assuming $\eta = 6e/16$.

The main result of this section is that the set of Λ^*-candidates is small. To put this size into context we note that the total number of entries considered for the Λ^*-decomposition rule is given by

$$\Omega(n) = \sum_{m=1}^{n} (n - m + 1).$$

Theorem 3. *Suppose an mfe g-structure over an interval of length m is irreducible with probability $\mathbf{d}^*_g(m)/\mathbf{d}_g(m)$, then the expected number of candidates of g-structures for sequences of lengths n satisfies*

$$\mathbb{E}_g(n) = \Theta(n^2)$$

and furthermore, setting $\overline{\mathbb{E}}_g(n) = \mathbb{E}_g(n)/\Omega(n)$ we have

$$\overline{\mathbb{E}}_g(n) \sim \mathbf{d}^*_g(n)/\mathbf{d}_g(n) \sim b_g,$$

where $b_g > 0$ is a constant.

We provide an illustration of Theorem 3 in Figure 9.

Proof. We proof the theorem by quantifying the probability of $[i,j]$ being a Λ^*-candidate. In this case any (not necessarily unique) sub-structure, realizing the optimal solution $L_{i,j}$, is Λ^*-irreducible, and therefore an irreducible structure over $[i,j]$.

Let $m = (j - i + 1)$, by assumption, the probability that $[i,j]$ is a candidate conditional to the existence of a substructure over $[i,j]$ is given by

$$\mathbb{P}_*([i,j] \mid [i,j] \text{ is a candidate }) = \frac{\mathbf{d}^*_g(m)}{\mathbf{d}_g(m)}, \tag{8}$$

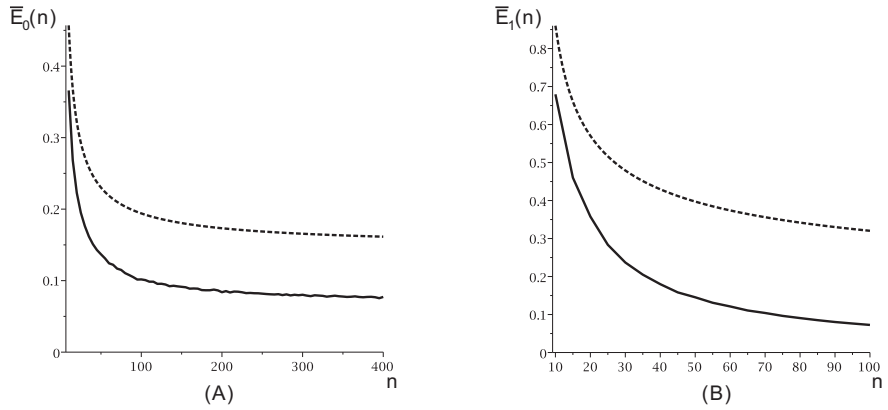

Figure 9: The expected number of candidates for secondary and 1-structures, $\overline{\mathbb{E}}_0(n)$ and $\overline{\mathbb{E}}_1(n)$: we compute the expected number of candidates obtained by folding 100 random sequences for secondary structures (A)(solid) and 1-structures (B)(solid). We also display the theoretical expectations implied by Theorem 3 (A)(dashed) and (B)(dashed).

Note that $\mathbb{P}_*([i,j] \mid [i,j]$ is a candidate) does not depend on the relative location of the interval but only on the interval-length. Let $\mathbb{P}_g(m) = \mathbf{d}_g^*(m)/\mathbf{d}_g(m)$, then according to Theorem 1,

$$\begin{aligned}(1-\epsilon)a_g m^{3(g-\frac{1}{2})}\gamma^m &\leq \mathbf{d}_g(m) \leq (1+\epsilon)a_g m^{3(g-\frac{1}{2})}\gamma^m,\\ (1-\epsilon)k_g m^{3(g-\frac{1}{2})}\gamma^m &\leq \mathbf{d}_g^*(m) \leq (1+\epsilon)k_g m^{3(g-\frac{1}{2})}\gamma^m,\end{aligned}$$

for $m \geq m_0$ where $m_0 > 0$ and $0 < \epsilon < 1$ are constants. On the one hand

$$\mathbb{P}_g(m) = \frac{\mathbf{d}_g^*(m)}{\mathbf{d}_g(m)} \leq \frac{(1+\epsilon)a_g m^{3(g-\frac{1}{2})}\gamma^m}{(1-\epsilon)k_g m^{3(g-\frac{1}{2})}\gamma^m} = (1+\epsilon')\frac{a_g}{k_g} = (1+\epsilon')b_g, \tag{9}$$

where $b_g = a_g/k_g > 0$ is a constant. On the other hand, we have

$$\mathbb{P}_g(m) = \frac{\mathbf{d}_g^*(m)}{\mathbf{d}_g(m)} \geq \frac{(1-\epsilon)a_g m^{3(g-\frac{1}{2})}\gamma^m}{(1+\epsilon)k_g m^{3(g-\frac{1}{2})}\gamma^m} = (1-\epsilon'')\frac{a_g}{k_g} = (1-\epsilon'')b_g. \tag{10}$$

Setting $\epsilon = \max\{\epsilon', \epsilon''\}$, we can conclude that $\mathbb{P}_g(m) \sim \mathbf{d}_g^*(m)/\mathbf{d}_g(m)$, see Fig. 10.

We next study the expected number of candidates over an interval of length m. To this end let

$$X_m = |\{[i,j] \mid [i,j] \text{ is a } \Lambda^*\text{-candidate of length } m\}|.$$

The expected cardinality of the set of Λ^*-candidates of length $m = (j-i+1)$ encountered in the DP-algorithm is given by

$$\mathbb{E}_g(X_m) \leq (n-(m-1))\,\mathbb{P}_g(m),$$

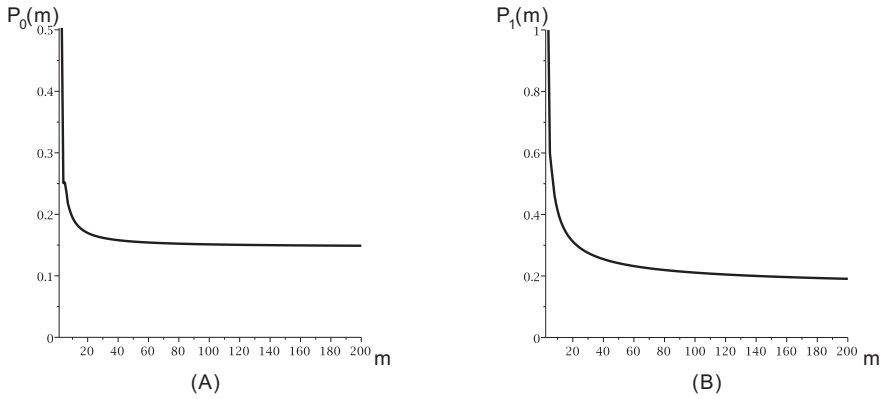

Figure 10: The probability distribution of $\mathbb{P}_0(m)$ (A) and $\mathbb{P}_1(m)$ (B).

since there are $n-(m-1)$ starting points for such an interval $[i,j]$. Therefore, by linearity of expectation, for sufficiently large $m > m_0$, $\mathbb{P}_g(m) \leq (1+\epsilon)b_g$ with ϵ being a small constant. Thus we have

$$\mathbb{E}_g(n) = \mathbb{E}_g(\sum_m X_m) \leq \sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1+\epsilon)b_g \sum_{m=m_0}^{n}(n-m+1). \tag{11}$$

Consequently, the expected size of the Λ^*-candidate set is $\Theta(n^2)$. We proceed by comparing the expected number of candidates of a sequence with length n with $\Omega(n)$,

$$\begin{aligned}
\frac{\mathbb{E}_g(n)}{\Omega(n)} &\leq \frac{\sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1+\epsilon)b_g\sum_{m=m_0}^{n}(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \\
&\leq (1+\epsilon)b_g + \frac{\sum_{m=1}^{m_0}(\mathbb{P}_g(m) - (1+\epsilon)b_g)(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \\
&\leq (1+\epsilon)b_g + \frac{k\cdot n}{n^2}.
\end{aligned}$$

For sufficient large $n \geq n_0$, $\mathbb{E}_g(n)/\Omega(n) \leq (1+\epsilon')b_g$. Furthermore

$$\frac{\mathbb{E}_g(n)}{\Omega(n)} \geq \frac{\sum_{m=1}^{m_0}(n-m+1)\mathbb{P}_g(m) + (1-\epsilon)b_g\sum_{m=m_0}^{n}(n-m+1)}{\sum_{m=1}^{n}(n-m+1)} \geq (1-\epsilon)b_g,$$

from which we can conclude $\mathbb{E}_g(n)/\Omega(n) \sim \mathbf{d}_g^*(m)/\mathbf{d}_g(m) \sim b_g$ and the theorem is proved. □

Loop-based energies

In this section we discuss the more realistic loop-based energy model of RNA secondary structure folding. To be precise we evoke here instead of two trivariate GFs $\mathbf{F}(z,t,v)$ and $\mathbf{F}^*(z,t,v)$ counting secondary structures over n vertices that filter energy and arcs.

This becomes necessary since the loop-based model distinguishes between arcs and energy. The "cancelation" effect or reparameterization of stickiness [33] to which we referred to before does not appear in this

context. Thus we need both an arc- as well as an energy-filtration.

A further complication emerges. In difference to the GFs $\mathbf{E}_g(z,t)$ and $\mathbf{E}_g^*(z,t)$ the new GFs are not simply obtained by formally substituting $(tz^2/((tz^2-z+1)^2)$ into the power series $\mathbf{D}_g(z)$ and $\mathbf{D}_g^*(z)$ as bivariate terms. The more complicated energy model requires a specific recursion for irreducible secondary structures.

The energy model used in prediction secondary structure is more complicated than the simple arc-based energy model. Loops which are formed by arcs as well as isolated vertexes between the arcs are considered to give energy contribution. Loops are categorized as hairpin loops (no nested arcs), interior loops (including bulge loops and stacks) and multi-loops (more than two arc nested), see Figure 11. An arbitrary secondary structure can be uniquely decomposed into a collection of mutually disjoint loops. A result of the particular energy parameters [3] is that the energy model prefers interior loops, in particular stacks (no isolated vertex between two parallel arc), and disfavors multi-loops. Base on this observation, we give a simplified energy model for a loop λ contained in secondary structure by

- $f(\lambda) = -0.5$ if ℓ is a hairpin loop,
- $f(\lambda) = 1$ if ℓ is an interior loop,
- $f(\lambda) = -5$ if ℓ is a multi-loop,

where λ is a loop. The weight for a secondary structure δ accordingly is given by

$$f(\delta) = \sum_{\lambda \in \delta} f(\lambda). \tag{12}$$

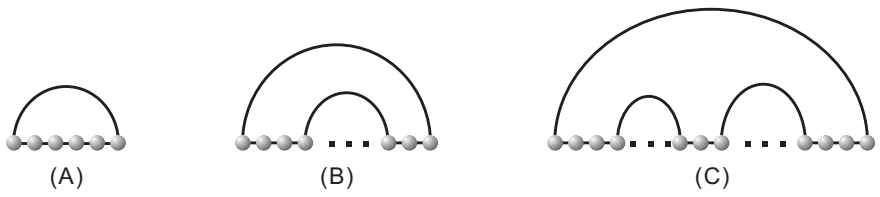

Figure 11: Diagram representation of loop types: (A) hairpin loop, (B) interior loop, (C) multi-loop.

Let $\mathbf{F}_0^*(z)$ and $\mathbf{F}_0(z)$ be the GFs obtained by setting $t = e$ and $v = 6/16$ in $\mathbf{F}^*(z,t,v)$ and $\mathbf{F}(z,t,v)$, where e is the Euler number. This means we find a suitable parameterization which brings us back to a simple univariate GF.

Lemma 3. *The weight function of RNA secondary structures,* $\mathbf{F}_0^*(z)$*, satisfies*

$$\mathbf{F}_0^*(z) = \frac{6}{16}e^{0.5}z^2\frac{z}{1-z} + \frac{6}{16}e^1z^2\left(\frac{1}{1-z}\right)^2\mathbf{F}_0^*(z) + \frac{6}{16}e^{-5}z^2\frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}. \tag{13}$$

and $\mathbf{F}^*(z)$ *is uniquely determined by the above equation. Furthermore*

$$\mathbf{F}_0(z) = \frac{1}{1-z}\frac{1}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}. \tag{14}$$

Proof. We first consider the GF $\mathbf{F}_0^*(z)$ whose coefficient of z^n denotes the total weight of irreducible secondary structures over n vertexes, where $(1,n)$ is an arc. Thus it gives a term $6/16z^2$. Isolated vertex lead to the term

$$z^p\sum_{i=0}^{\infty}z^i = z^p\frac{1}{1-z},$$

where p denotes the minimum number of isolated vertexes to be inserted. Depending on the types of loops formed by (i,n), we have

- hairpin loops: $\frac{z}{1-z}$,
- interior loops: $\mathbf{F}_0^*(z)\left(\frac{1}{1-z}\right)^2$,
- multi-loops: there are at least two irreducible substructures, as well as isolated vertices, thus

$$\frac{1}{1-z}\sum_{i=2}^{\infty}\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^i = \frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}.$$

We compute

$$\mathbf{F}_0^*(z) = \frac{6}{16}\left(e^{0.5}z^2\frac{z}{1-z} + e^1z^2\left(\frac{1}{1-z}\right)^2\mathbf{F}_0^*(z) + e^{-5}z^2\frac{\left(\mathbf{F}_0^*(z)\frac{1}{1-z}\right)^2}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}\frac{1}{1-z}\right),$$

which establishes the recursion. The uniqueness of the solution as a power series follows from the fact that each coefficient can evidently be recursively computed.

An arbitrary secondary structure can be considered as a sequence of irreducible substructure with certain intervals of isolated vertexes. Thus

$$\mathbf{F}_0(z) = \frac{1}{1-z}\sum_{i=0}^{\infty}\frac{1}{1-z}\mathbf{F}_0^*(z) = \frac{1}{1-z}\frac{1}{1-\mathbf{F}_0^*(z)\frac{1}{1-z}}.$$

□

Lemma 4. $\mathbf{F}_0^*(z)$ *and* $\mathbf{F}_0(z)$ *have the same singular expansion.*

$$\mathbf{f}_0^*(n) \sim \alpha n^{-\frac{3}{2}}\gamma^n, \quad \textit{and} \quad \mathbf{f}_0(n) \sim \beta n^{-\frac{3}{2}}\gamma^n, \tag{15}$$

where $\alpha \approx 0.24$ *and* $\beta \approx 2.88$ *are constants and* $\gamma \approx 2.1673$

Proof. Solving eq. 13 we obtain a unique solution for $\mathbf{F}_0^*(z)$ whose coefficient are all positive. Observing the dominant singularity of $\mathbf{F}_0^*(z)$ it is $\rho \approx 0.4614$. $\mathbf{F}_0(z)$ is a function of $\mathbf{F}_0^*(z)$ and we examine the real root of minimal modulus of $1 - \mathbf{F}_0^*(z)\frac{1}{1-z} = 0$ is bigger than ρ. Then by the supercritical paradigm [41] applying, $\mathbf{F}_0(z)$ and $\mathbf{F}_0^*(z)$ have identical exponential growth rates. Furthermore, $\mathbf{F}_0^*(z)$ and $\mathbf{F}_0(z)$ have the same sub-exponential factor $n^{-\frac{3}{2}}$, hence the lemma. □

Theorem 4. *Suppose an mfe secondary structure over an interval of length* m *is irreducible with probability* $\mathbb{P}_0(m) = \frac{\mathbf{f}_0^*(m)}{\mathbf{f}_0(m)}$*, then the expected number of candidates for sequences of lengths* n *is*

$$\mathbb{E}_0(n) = \Theta(n^2)$$

and furthermore, setting $\overline{\mathbb{E}}_g(n) = \mathbb{E}_g(n)/\Omega(n)$*, we have*

$$\overline{\mathbb{E}}_0(n) \sim \mathbf{f}_0^*(n)/\mathbf{f}_0(n) \sim b,$$

where $b = \alpha/\beta \approx 0.08$.

Proof. By Lemma 4 we have $\mathbf{f}_0^*(m)/\mathbf{f}_0(m) \sim b$ where b is a constant. The proof is completely analogous to that of Theorem 3. □

We show the distribution of $\mathbb{P}_0(m)$ and $\overline{\mathbb{E}}_0(n)$ in Figure 12.

Results and Discussion

In this paper we quantify the effect of sparsification of the particular decomposition rule Λ^*. This rule splits and interval and thereby separates concatenated substructures. The sparsification of Λ^* alone is claimed to provide a speed up of up to a linear factor of the DP-folding of RNA secondary structures [24]. A similar conclusion is drawn in [26] where the sparsification of RNA-RNA interaction structures is shown to experience also a linear reduction in time complexity. Both papers [24, 26] base their conclusion on the validity of the polymer-zeta property discussed in Section **Sparsification**.

For the folding of pseudoknot structures there may however exist non-sparsifiable rules in which case the overall time complexity is not reduced. The key object here is the *set of candidates* and we provide an

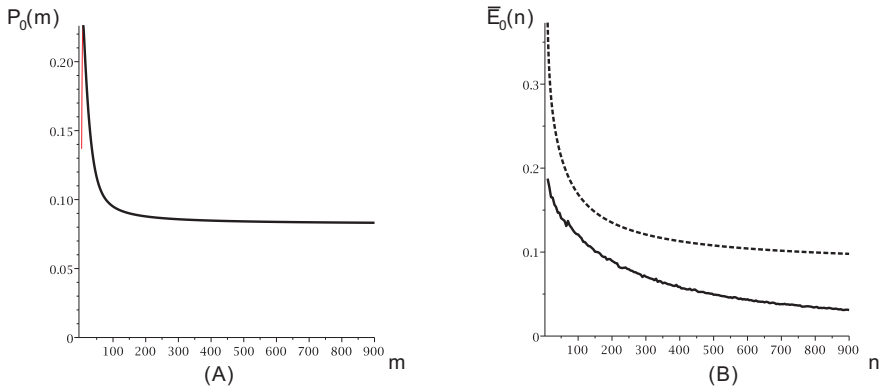

Figure 12: The distribution of $\mathbb{P}_0(m)$ (A) and $\overline{\mathbb{E}}_0(n)$ obtained by folding 100 random sequences on the loop-based model (B)(solid), as well as the theoretical expectation implied by Theorem 4 (B)(dashed).

analysis of Λ^*-candidates by combinatorial means. In general, the connection between candidates, i.e. unions of disjoint intervals and the combinatorics of structures is actually established by the algorithm itself via backtracking: at the end of the DP-algorithm a structure is being generated that realizes the previously computed energy as mfe-structure. This connects intervals and sub-structures.

So, does polymer-zeta apply in the context of RNA structures? In fact polymer-zeta would follow *if* the intervals in question are distributed as in uniformly sampled structures. This however, is far from reasonable, due to the fact that the mfe-algorithm deliberately designs some mfe structure over the given interval. What the algorithm produces is in fact antagonistic to uniform sampling. We here wish to acknowledge the help of one anonymous referee in clarifying this point.

Our results clearly show that the polymer-zeta property, i.e. the probability of an irreducible structure over an interval of length m satisfies a formula of the form

$$\mathbb{P}(\text{there exists an irreducible structure over } [1,m]) = b\, m^{1+c}, \quad \text{where } \; b, c > 0. \tag{16}$$

does not apply for RNA structures. The theoretical findings from self-avoiding walks [30] unfortunately do not allow to quantify the expected number of candidates of the Λ^*-rule in RNA folding.

That the polymer-zeta property does not hold for RNA has also been observed in the context of the limit distribution of the 5'-3' distances of RNA secondary structures [42]. Here it is observed that long arcs, to be precise arcs of lengths $O(n)$ *always* exist. This is of course a contradiction to eq. (16).

The key to quantification of the expected number of candidates is the singularity analysis of a pair of energy-filtered GF, namely that of a class of structures and that of the subclass of all such structures that are irreducible. We show that for various energy models the singular expansions of both these functions

are essentially *equal*–modulo some constant. This implies that the expected number of candidates is $\Theta(n^2)$ and all constants can explicitly be computed from a detailed singularity analysis. The good news is that depending on the energy model, a significant constant reduction, around 95% can be obtained. This is in accordance with data produced in [25] for the mfe-folding of random sequences. There a reduction by 98% is reported for sequences of length ≥ 500.

Our findings are of relevance for numerous results, that are formulated in terms of sizes of candidate sets [27]. These can now be quantified. It is certainly of interest to devise a full fledged analysis of the loop-based energy model. While these computations are far from easy our framework shows how to perform such an analysis.

Using the paradigm of gap-matrices Backofen has shown [27] that the sparsification of the DP-folding of RNA pseudoknot structures exhibits additional instances, where sparsification can be applied, see Fig. 6 (B). Our results show that the expected number of candidates is $\Theta(n^2)$, where the constant reduction is around 90%. This is in fact very good new since the sequence length in the context of RNA pseudoknot structure folding is in the order of hundreds of nucleotides. So sparsification of further instances does have an significant impact on the time complexity of the folding.

Proofs

In this section, we prove Lemma 2 and Theorem 2.

Proof for Lemma 2: let $\mathbf{D}(z,u)$ and $\mathbf{D}^*(z,u)$ be the bivariate GF $\mathbf{D}(z,u) = \sum_{n\geq 0}\sum_{g=0}^{\lfloor \frac{n}{2}\rfloor} \mathbf{d}_g(n) z^n u^g$, and $\mathbf{D}^*(z,u) = \sum_{n\geq 1}\sum_{g=0}^{\lfloor \frac{n}{2}\rfloor} \mathbf{d}^*_g(n) z^n u^g$. Suppose a structure contains exactly j irreducible structures, then

$$\mathbf{D}(z,u) = \sum_{j\geq 0} \mathbf{R}(z,u)^j = \frac{1}{1-\mathbf{R}(z,u)} \tag{17}$$

and

$$\mathbf{D}^*_g(z) = [u^g]\mathbf{D}^*(z,u) = -[u^g]\frac{1}{\mathbf{D}(z,u)}, \quad g \geq 1, \tag{18}$$

as well as $\mathbf{D}^*_0(z) = 1 - [u^0]\frac{1}{\mathbf{D}(z,u)}$. Let $\mathbf{F}(z,u) = \sum_{n\geq 0}\sum_{g\geq 0} \mathbf{f}_g(n) z^n u^g = \frac{1}{\mathbf{D}(z,u)}$. Then $\mathbf{F}(z,u)\mathbf{D}(z,u) = 1$, whence for $g \geq 1$,

$$\sum_{g_1=0}^{g} \mathbf{F}_{g_1}(z)\mathbf{D}_{g-g_1}(z) = [u^g]\mathbf{F}(z,u)\mathbf{D}(z,u) = 0, \tag{19}$$

and $\mathbf{F}_0(z)\mathbf{D}_0(z) = 1$, where $\mathbf{F}_g(z) = \sum_{n\geq 0}\mathbf{f}_g(n)z^n = [u^g]\mathbf{F}(z,u) = [u^g]\frac{1}{\mathbf{D}(z,u)}$. Furthermore, we have $\mathbf{F}_0(z) = \frac{1}{\mathbf{D}_0(z)}$ and

$$\mathbf{F}_g(z) = -\frac{\sum_{g_1=0}^{g-1}\mathbf{F}_{g_1}(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}, \quad g \geq 1, \tag{20}$$

which implies $\mathbf{D}_0^*(z) = 1 - \mathbf{F}_0(z) = 1 - \frac{1}{\mathbf{D}_0(z)}$ and

$$\mathbf{D}_g^*(z) = -\mathbf{F}_g(z) = -\frac{(\mathbf{D}_0^*(z)-1)\mathbf{D}_g(z) + \sum_{g_1=1}^{g-1} \mathbf{D}_{g_1}^*(z)\mathbf{D}_{g-g_1}(z)}{\mathbf{D}_0(z)}. \tag{21}$$

Proof for Theorem 2 Let $[n]_k$ denotes the set of compositions of n having k parts, i.e. for$\sigma \in [n]_k$ we have $\sigma = (\sigma_1, \ldots, \sigma_k)$ and $\sum_{i=1}^k \sigma_i = n$.

Claim.

$$\mathbf{D}_{g+1}^*(z) = \frac{\mathbf{D}_{g+1}(z)}{\mathbf{D}_0(z)^2} + \sum_{j=0}^{g-1} \frac{(-1)^{g+2-j}}{\mathbf{D}_0(z)^{g+2-j}} \left(\sum_{\sigma \in [g+1]_{g+1-j}} \prod_{i=1}^{g+1-j} \mathbf{D}_{\sigma_i}(z) \right). \tag{22}$$

We shall prove the claim by induction on g. For $g = 1$ we have

$$\mathbf{D}_1^*(x) = \frac{\mathbf{D}_1(z)}{(\mathbf{D}_0(z))^2}, \tag{23}$$

whence eq. (22) holds for $g = 1$. By induction hypothesis, we may now assume that for $j \leq g$, eq. (22) holds. According to Lemma 2, we have

$$\begin{aligned}
\mathbf{D}_{g+1}^*(z) &= -\frac{(\mathbf{D}_0^*(z)-1)\mathbf{D}_{g+1}(z) + \sum_{g_1=1}^{g} \mathbf{D}_{g_1}^*(z)\mathbf{D}_{g+1-g_1}(z)}{\mathbf{D}_0(z)} \\
&= \frac{\mathbf{D}_{g+1}(z)}{\mathbf{D}_0(z)^2} - \sum_{g_1=1}^{g} \left(\frac{\mathbf{D}_{g_1}(z)}{\mathbf{D}_0(z)^3} + \sum_{j=0}^{g_1-2} \frac{(-1)^{g_1+1-j}}{\mathbf{D}_0(z)^{g_1+2-j}} \left(\sum_{\sigma \in [g_1]_{g_1-j}} \prod_{i=1}^{g_1-j} \mathbf{D}_{\sigma_i}(z) \right) \right) \mathbf{D}_{g+1-g_1}(z).
\end{aligned}$$

We next observe

$$-\sum_{g_1=1}^{g} \frac{\mathbf{D}_{g_1}(z)}{\mathbf{D}_0(z)^3} \mathbf{D}_{g+1-g_1}(z) = \frac{(-1)^{g+2-(g-1)}}{\mathbf{D}_0(z)^{g+2-(g-1)}} \left(\sum_{\sigma' \in [g+1]_{g+1-(g-1)}} \prod_{i=1}^{g+1-(g-1)} \mathbf{D}_{\sigma'_i}(z) \right), \tag{24}$$

and setting $h = g_1 - j$ we obtain,

$$\begin{aligned}
&-\sum_{g_1=1}^{g} \sum_{j=0}^{g_1-2} \frac{(-1)^{g_1+1-j}}{\mathbf{D}_0(z)^{g_1+2-j}} \left(\sum_{\sigma \in [g_1]_{g_1-j}} \prod_{i=1}^{g_1-j} \mathbf{D}_{\sigma_i}(z) \right) \mathbf{D}_{g+1-g_1}(z) \\
&= \sum_{g_1=1}^{g} \sum_{h=2}^{g_1} \frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}} \left(\sum_{\sigma \in [g_1]_h} \prod_{i=1}^{h} \mathbf{D}_{\sigma_i}(z) \right) \mathbf{D}_{g+1-g_1}(z) \\
&= \sum_{h=2}^{g} \frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}} \left(\sum_{g_1=h}^{g} \left(\sum_{\sigma \in [g_1]_h} \prod_{i=1}^{h} \mathbf{D}_{\sigma_i}(z) \right) \mathbf{D}_{g+1-g_1}(z) \right) \\
&= \sum_{h=2}^{g} \frac{(-1)^{h+2}}{\mathbf{D}_0(z)^{h+2}} \left(\sum_{\sigma' \in [g+1]_{h+1}} \prod_{i=1}^{h+1} \mathbf{D}_{\sigma'_i}(z) \right)
\end{aligned}$$

and setting $j = g - h$

$$= \sum_{j=0}^{g-2} \frac{(-1)^{g+2-j}}{\mathbf{D}_0(z)^{g+2-j}} \left(\sum_{\sigma' \in [g+1]_{g+1-j}} \prod_{i=1}^{g+1-j} \mathbf{D}_{\sigma'_i}(z) \right).$$

Consequently, the Claim holds for any $g \geq 1$.

For any $g \geq 1$, we have [23]

$$\mathbf{D}_g(z) = \frac{1}{z^2-z+1}\frac{\mathbf{P}_g(u)}{(1-4u)^{3g-1/2}}, \qquad \mathbf{D}_0(z) = \frac{1}{z^2-z+1}\frac{2}{(1+\sqrt{1-4u})},$$

where $\mathbf{P}_g(u)$ is a polynomial with integral coefficients of degree at most $(3g-1)$, $\mathbf{P}_g(1/4) \neq 0$, $[u^{2g}]\mathbf{P}_g(u) \neq 0$ and $[u^h]\mathbf{P}_g(u) = 0$ for $0 \leq h \leq 2g-1$. Let $u = \frac{z^2}{(z^2-z+1)^2}$, the Claim provides in this context the following interpretation of $\mathbf{D}_g^*(z)$

$$\frac{1}{z^2-z+1}\mathbf{D}_g^*(z) = \frac{\mathbf{P}_g(u)}{(1-4u)^{3g-1/2}}\left(\frac{1+\sqrt{1-4u}}{2}\right)^2 + \sum_{j=0}^{g-2}\left(-\frac{1+\sqrt{1-4u}}{2}\right)^{g+1-j}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j}{2}}}, \tag{25}$$

and

$$\begin{aligned}
&\sum_{j=0}^{g-2}\left(-\frac{1+\sqrt{1-4u}}{2}\right)^{g+1-j}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j}{2}}} \\
&= \sum_{j=0}^{g-2}\sum_{k=0}^{g+1-j}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{k}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{g-j+k}{2}}} \\
&= \sum_{j=0}^{g-2}\sum_{s=g-j}^{2g+1-2j}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\frac{\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)}{(1-4u)^{3g-\frac{s}{2}}}.
\end{aligned}$$

As $0 \leq j \leq g-2$ and $g-j \leq s \leq 2g+1-2j$, we have $s \geq 2$. Consequently we arrive at

$$\frac{1}{z^2-z+1}\mathbf{D}_g^*(z) = \frac{\mathbf{U}_g(u)}{(1-4u)^{3g-1/2}} + \frac{\mathbf{V}_g(u)}{(1-4u)^{3g-1}}, \tag{26}$$

where

$$\begin{aligned}
\mathbf{U}_g(u) &= \frac{\mathbf{P}_g(u)}{4} + \frac{\mathbf{P}_g(u)(1-4u)}{4} \\
&+ \sum_{j=0}^{g-2}\sum_{\substack{g-j\leq s\leq 2g+1-2j \\ s \text{ is odd}}}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\left(\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)\right)(1-4u)^{\frac{s-1}{2}},
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{V}_g(u) &= \frac{\mathbf{P}_g(u)}{2} + \left(-\frac{1}{2}\right)^3\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(u)\right) + 3\left(-\frac{1}{2}\right)^3\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(u)\right)(1-4u) \\
&+ \sum_{j=0}^{g-3}\sum_{\substack{g-j\leq s\leq 2g+1-2j \\ s \text{ is even}}}\left(-\frac{1}{2}\right)^{g+1-j}\binom{g+1-j}{s-g+j}\left(\sum_{\sigma\in[g]_{g-j}}\prod_{i=1}^{g-j}\mathbf{P}_{\sigma_i}(u)\right)(1-4u)^{\frac{s-2}{2}}.
\end{aligned}$$

We have for $\sigma \in [g]_k$, $k \geq 1$

$$[u^h]\left(\sum_{\sigma\in[g]_k}\prod_{i=1}^{k}\mathbf{P}_{\sigma_i}(u)\right) = \sum_{\sigma\in[g]_k}\prod_{i=1}^{k}[u^{h_i}]\mathbf{P}_{\sigma_i}(u),$$

where $\sum_{i=1}^{k} h_i = h$, $h_i \geq 0$. Then we obtain that

$$[u^h]\left(\sum_{\sigma\in[g]_k}\prod_{i=1}^{k}\mathbf{P}_{\sigma_i}(u)\right) = 0, \quad 0 \leq h \leq 2g-1. \tag{27}$$

Since $[u^{h_i}]\mathbf{P}_{\sigma_i}(u) = 0$, $h_i \leq 2\sigma_i - 1$, $[u^{2\sigma_i}]\mathbf{P}_{\sigma_i}(u) \neq 0$ and $\sum_{i=1}^{k}\sigma_i = g$. Thus for $0 \leq h \leq 2g-1$,

$$[u^h]\mathbf{U}_g(u) = 0 \quad \text{and} \quad [u^h]\mathbf{V}_g(u) = 0. \tag{28}$$

As shown in [23] we have

$$\mathbf{P}_g(1/4) = \frac{\Gamma\left(g-1/6\right)\Gamma\left(g+1/2\right)\Gamma\left(g+1/6\right)9^g 4^{-g}}{6\pi^{3/2}\Gamma\left(g+1\right)} \tag{29}$$

and we obtain $\mathbf{U}_g(1/4) = \mathbf{P}_g(1/4)/4$. Furthermore,

$$\mathbf{V}_g(1/4) = \frac{\mathbf{P}_g(1/4)}{2} + \left(-\frac{1}{2}\right)^3\left(\sum_{\sigma\in[g]_2}\prod_{i=1}^{2}\mathbf{P}_{\sigma_i}(1/4)\right) = \frac{1}{8}\left(4\mathbf{P}_g(1/4) - \sum_{j=1}^{g-1}\mathbf{P}_j(1/4)\mathbf{P}_{g-j}(1/4)\right) \neq 0.$$

We can recruit the computation of [23] in order to observe $4\mathbf{P}_g(1/4) - \sum_{j=1}^{g-1}\mathbf{P}_j(1/4)\mathbf{P}_{g-j}(1/4) \neq 0$. In order to compute the bivariate GF, $\mathbf{E}^*_g(z,t)$, we only need to replace in eq. (22) $\mathbf{D}_g(z)$ by $\mathbf{E}_g(z,t)$ and the proof is completely analogous.

Author's contributions

Fenix W.D. Huang and Christian M. Reidys contributed equally to research and manuscript.

Acknowledgements

We want to thank Thomas J.X. Li for discussions and comments. We want to thank an anonymous referee for pointing out an incorrect assumption of first version of this paper. His comments have led to a much improved version of the paper.

References

1. Bailor MH, Sun X, Al-Hashimi HM: **Topology Links RNA Secondary Structure with Global Conformation, Dynamics, and Adaptation**. *Science* 2010, **327**:202–206.
2. Tabaska JE, Cary RB, Gabow HN, Stormo GD: **An RNA folding method capable of identifying pseudoknots and base triples**. *Bioinformatics* 1998, **14**:691–699.
3. Mathews D, Sabina J, Zuker M, Turner D: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure**. *J. Mol. Biol.* 1999, **288**:911–940.
4. Smith T, Waterman M: **RNA secondary structure**. *Math. Biol.* 1978, **42**:31–49.
5. Zuker M: **On finding all suboptimal foldings of an RNA molecule**. *Science* 1989, **244**:48–52.
6. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures**. *Monatsh. Chem.* 1994, **125**:167–188.
7. Rivas E, Eddy SR: **A Dynamic Programming Algorithm for RNA Structure Prediction Including Pseudoknots**. *J. Mol. Biol.* 1999, **285**:2053–2068.
8. Uemura Y A Hasegawa, Kobayashi S, Yokomori T: **Tree adjoining grammars for RNA structure prediction**. *Theor. Comp. Sci.* 1999, **210**:277–303.
9. Akutsu T: **Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots**. *Discr. Appl. Math.* 2000, **104**:45–62.
10. Lyngsø RB, Pedersen CN: **RNA pseudoknot prediction in energy-based models**. *J. Comp. Biol.* 2000, **7**:409–427.
11. Cai L, Malmberg RL, Wu Y: **Stochastic modeling of RNA pseudoknotted structures: a grammatical approach**. *Bioinformatics* 2003, **19 S1**:i66–i73.
12. Dirks RM, Pierce NA: **A partition function algorithm for nucleic acid secondary structure including pseudoknots**. *J. Comput. Chem.* 2003, **24**:1664–1677.
13. Deogun JS, Donis R, Komina O, Ma F: **RNA secondary structure prediction with simple pseudoknots**. In *Proceedings of the second conference on Asia-Pacific bioinformatics (APBC 2004)*, Australian Computer Society 2004:239–246.
14. Reeder J, Giegerich R: **Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics**. *BMC Bioinformatics* 2004, **5**:104.
15. Li H, Zhu D: **A New Pseudoknots Folding Algorithm for RNA Structure Prediction**. In *COCOON 2005, Volume 3595*. Edited by Wang L, Berlin: Springer 2005:94–103.
16. Matsui H, Sato K, Sakakibara Y: **Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures**. *Bioinformatics* 2005, **21**:2611–2617.
17. Kato Y, Seki H, Kasami T: **RNA Pseudoknotted Structure Prediction Using Stochastic Multiple Context-Free Grammar**. *IPSJ Digital Courier* 2006, **2**:655–664.
18. Chen HL, Condon A, Jabbari H: **An $O(n^5)$ Algorithm for MFE Prediction of Kissing Hairpins and 4-Chains in Nucleic Acids**. *J. Comp. Biol.* 2009, **16**:803–815.
19. Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, Nebel ME: **Topology and prediction of RNA pseudoknots**. *Bioinformatics* 2011, **27**:1076–1085.
20. Waterman MS: **Secondary structure of single-stranded nucleic acids**. *Adv. Math. (Suppl. Studies)* 1978, **1**:167–212.
21. Orland H, Zee A: **RNA folding and large N matrix theory**. *Nuclear Physics B* 2002, **620**:456–476.
22. Bon M, Vernizzi G, Orland H, Zee A: **Topological Classification of RNA Structures**. *J. Mol. Biol.* 2008, **379**:900–911.
23. Andersen JE, Penner RC, Reidys CM, Waterman MS: **Topological classification and enumeration of RNA structrues by genus**. *J. Math. Biol.* 2011. [Submitted].
24. Wexler Y, Zilberstein C, Ziv-Ukelson M: **A study of accessible motifs and RNA complexity**. *J. Comput. Biol.* 2007, **14(6)**:856–872.

25. Backofen R, Tsur D, Zakov S, Ziv-Ukelson M: **Sparse RNA folding: Time and space efficient algorithms**. *J. Disc. Algor.* 2011, **9(1)**:12–31.

26. Salari R, Möhl M, Will S, Sahinalp C, Backofen R: **Time and space efficient RNA-RNA interaction prediction via sparse folding**. *Proc. of RECOMB* 2010, **6044**:473–490.

27. Möhl M, Salari R, Will S, Backofen R, Sahinalp SC: **Sparsification of RNA structure prediction including pseudoknots**. *Algorithms Mol. Biol.* 2010, **5**:39.

28. McCaskill JS: **The equilibrium partition function and base pair binding probabilities for RNA secondary structure**. *Biopolymers* 1990, **29**:1105–1119.

29. Kafri Y, Mukamel D, Peliti L: **Why is the DNA Denaturation Transition First Order?** *Phys. Rev. Lett.* 2000, **85**:4988–4991.

30. Kabakcioglu A, Stella AL: **A scale-free network hidden in the collapsing polymer**. *Phys. .Rev. E* 2005, **72**:055102.

31. Vanderzande C: *Lattic models of polymers*. Cambridge University Press New York 1998.

32. **NCBI database**[http://www.ncbi.nlm.nih.gov/guide/dna-rna/#downloads_].

33. Nebel ME: **Investigation of the Bernoulli model for RNA secondary structures**. *Bull. math. biol.* 2003, **66(5)**:925–964.

34. Zagier D: **On the distribution of the number of cycles of elements in symmetric groups**. *Nieuw Arch. Wisk. IV* 1995, **13**:489–495.

35. Loebl M, Moffatt I: **The chromatic polynomial of fatgraphs and its categorification**. *Adv. Math.* 2008, **217**:1558–1587.

36. Penner RC, Knudsen M, Wiuf C, Andersen JE: **Fatgraph models of proteins**. *Comm. Pure Appl. Math.* 2010, **63**:1249–1297.

37. Massey WS: *Algebraic Topology: An Introduction*. Springer-Veriag, New York 1967.

38. Penner RC, Waterman MS: **Spaces of RNA secondary structures**. *Adv. Math.* 1993, **101**:31–49.

39. Penner RC: **Cell decomposition and compactification of Riemann's moduli space in decorated Teichmüller theory**. In *Woods Hole Mathematics-perspectives in math and physics*. Edited by Tongring N, Penner RC, Singapore: World Scientific 2004:263–301. [ArXiv: math. GT/0306190].

40. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matching**. *SIAM J. Appl. Math.* 1978, **35**:68–82.

41. Flajolet P, Sedgewick R: *Analytic Combinatorics*. Cambridge University Press New York 2009.

42. Han HSW, Reidys CM: **The 5'-3' distance of RNA secondary structures**. *J. Comput. Biol.* [Submitted, e-print arXiv:1104.3316v2].

Figures

Figure 1 - RNA structures as planar graphs and diagrams

(A) an RNA secondary structure and (B) an RNA pseudoknot structure.

Figure 2 - Sparsification of secondary structure folding

Suppose the optimal solution $L_{i,j}$ is obtained from the optimal solutions $L_{i,k}$, $L_{k+1,q}$ and $L_{q+1,j}$. Based on the recursions of the secondary structures, $L_{i,k}$ and $L_{k+1,q}$ produce an optimal solution of $L_{i,q}$. Similarly, $L_{k+1,q}$ and $L_{q+1,j}$ produce an optimal solution of $L_{k+1,j}$. Now, in order to obtain an optimal solution of $L_{i,j}$ it is sufficient to consider either the grouping $L_{i,q}$ and $L_{q+1,j}$ or $L_{i,k}$ and $L_{k+1,j}$.

Figure 3 - What sparsification can and cannot prune

What sparsification can and cannot prune: (A) and (B) are two computation paths yielding the same optimal solution. Sparsification reduces the computation to path (A) where S_{i,k_1} is irreducible. (C) is another computation path with distinct leftmost irreducible over a different interval, hence representing a new candidate that cannot be reduced to (A) by the sparsification.

Figure 4 - Sparsification

Sparsification: L_v is alternatively realized via L_{v_1} and $L_{v'_2}$, or $L_{v'_1}$ and L_{v_3}. Thus it is sufficient to only consider one of the computation paths.

Figure 5 - The recursion solving the optimal solution for secondary structures

The recursion solving the optimal solution for secondary structures.

Figure 6 - Decomposition rules for pseudoknot structures of fixed genus

(A) three decompositions via the rule Λ^*, which is s-compatible to itself. We show that for Λ^* we obtain a linear reduction in time complexity. (B) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are s-compatible to Λ_1. A quantification of the candidate set is not implied by the polymer-zeta property. (C) three decomposition rules $\Lambda_1, \Lambda_2, \Lambda_3$ where Λ_2, Λ_3 are not s-compatible to Λ_1.

Figure 7 - RNA structures and diagram representation

A diagram over $\{1, \ldots, 40\}$. The arcs $(1, 21)$ and $(11, 33)$ are crossing and the dashed arc $(9, 10)$ is a 1-arc which is not allowed. This structure contains 3 stacks with length 7, 4 and 6, from left to right respectively.

Figure 8 - Irreducibility relative to a decomposition rule

the rule Λ^* splitting $S_{i,j}$ to $S_{i,k}$ and $S_{k+1,j}$, $S_{1,40}$ is not Λ^*-irreducible, while $S_{1,25}$ and $S_{28,40}$ are. However, for the decomposition rule Λ_2, which removes the outmost arc, $S_{28,40}$ is not Λ_2-irreducible while $S_{1,25}$ is.

Figure9 -The expected number of candidates for secondary and 1-structures $\overline{\mathbb{E}}_0(n)$ and $\overline{\mathbb{E}}_1(n)$

we compute the expected number of candidates obtained by folding 100 random sequences for secondary structures (A)(solid) and 1-structures (B)(solid). We also display the theoretical expectations implied by Theorem 3 (A)(dashed) and (B)(dashed).

Figure10- The probability distribution of $\mathbb{P}_0(m)$ and $\mathbb{P}_1(m)$

The probability distribution of $\mathbb{P}_0(m)$ (A) and $\mathbb{P}_1(m)$ (B)

Figure11 -Diagram representation of loop types

(A) hairpin loop, (B) interior loop, (C) multi-loop.

Figure12 -The distribution of $\mathbb{P}_0(m)$ (A) and $\overline{\mathbb{E}}_0(n)$

The distribution of $\mathbb{P}_0(m)$ (A) and $\overline{\mathbb{E}}_0(n)$ obtained by folding 100 random sequences on the loop-based model (B)(solid), as well as the theoretical expectation implied by Theorem 4 (B)(dashed).